

## PATENT ABSTRACTS OF JAPAN

(11)Publication number : 04-228153

(43)Date of publication of application : 18.08.1992

(51)Int.Cl.

G11B 20/18  
G11B 20/12

(21)Application number : 03-103708

(71)Applicant : INTERNATL BUSINESS MACH  
CORP <IBM>

(22)Date of filing : 14.03.1991

(72)Inventor : BEST JOHN S  
CHAINER TIMOTHY J  
GLASER THOMAS W  
GREENBERG RICHARD  
MUKHERJEE AVIJIT  
NEUBAUER JERRY L  
REIDENBACH JOHN R  
SCHOPP ROBERT E  
SCRANTON ROBERT A

(30)Priority

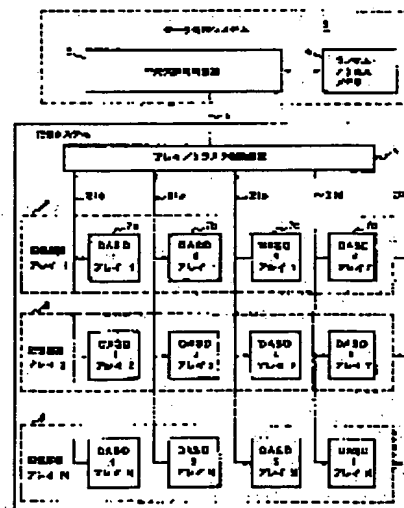
Priority number : 90 502215 Priority date : 30.03.1990 Priority country : US

## (54) DISK STORAGE DEVICE

(57)Abstract:

PURPOSE: To prevent the loss of data even when one driven is out of order by storing data in a first to a (n-1)th drives and parity information in an nth drive.

CONSTITUTION: A disk storage system includes an array/cluster controller 5 and plural storage mechanism arrays 7-9. The constitution of four (n=4) separated disk files 7a-7d into which the respective arrays are divided is preferable since a half/high disk drive is geometrically matched with an existing shape factor of a personal computer. The array is designed as the subsystem of a direct access storage device, is a system so that the data are stored and fetched at high speed and with high reliability and the data are dispersed between the constituent elements of the system.



## LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than

the examiner's decision of rejection or  
application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision  
of rejection]

[Date of requesting appeal against examiner's  
decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2003 Japan Patent Office

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平4-228153

(43) 公開日 平成4年(1992)8月18日

(51) Int.Cl. <sup>5</sup>	識別記号	庁内整理番号	F I	技術表示箇所
G 1 1 B 20/18	1 0 2	9074-5D		
20/12		9074-5D		

審査請求 有 請求項の数11(全 21 頁)

(21) 出願番号 特願平3-103708

(22) 出願日 平成3年(1991)3月14日

(31) 優先権主張番号 5 0 2 2 1 5

(32) 優先日 1990年3月30日

(33) 優先権主張国 米国 (US)

(71) 出願人 390009531

インターナショナル・ビジネス・マシー  
ズ・コーポレーション

INTERNATIONAL BUSIN  
ESS MACHINES CORPO  
RATION

アメリカ合衆国10504、ニューヨーク州  
アーモンク (番地なし)

(72) 発明者 ジョン・ステュワート・ベスト

アメリカ合衆国カリフォルニア州サンノ  
ゼ、オアクレスト・ドライブ 6486番地

(74) 代理人 弁理士 頓宮 孝一 (外4名)

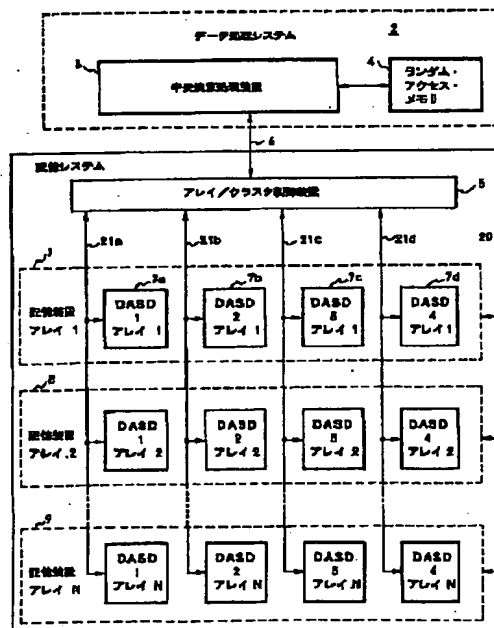
最終頁に続く

(54) 【発明の名称】 ディスク記憶システム

(57) 【要約】

【目的】 単一ドライブ・システムよりも信頼性の高い多重ドライブ・ディスク記憶システムを得る。

【構成】 N個のドライブのうち、第1ないし第N-1番目のドライブにデータを記憶させ、第N番目のドライブにパリティ情報を記憶させたハード・ディスク記憶システムが提供されるので1つのドライブに故障があったり、遮断されたりしてもデータが失われない。



## 【特許請求の範囲】

【請求項1】 (a) 磁気記録に適した磁性体のコーティングを有する基板から成るハード・ディスクと、

(b) 前記ディスクを回転させるためのスピンドル駆動手段と、

(c) データを読み出し、記録するために前記磁性体と作用結合するトランスジューサ手段と、

(d) 前記トランスジューサ手段を、前記ディスク上の複数のデータ・トラックのうちの選択された1本の上方で位置決めするためのアクチュエータ手段と、

(e) N個のドライブとを含み、データを前記ドライブのうち前記の第1のドライブないし第N-1番目のドライブに記憶させ、前記の第N番目のドライブにパリティ情報を記憶させる、ハード・ディスク記憶システム。

【請求項2】 N=4である、請求項1に記載のハード・ディスク記憶システム。

【請求項3】 さらに、ハード・ディスク記憶システムに記憶されるデータ・ブロックを受け取り、前記データ・ブロックをN-1個の部分と1個のパリティ部分に変換し、前記の各部分をそれぞれ前記ディスクの各1枚上に記憶するための制御手段を含む、請求項1に記載のハード・ディスク記憶システム。

【請求項4】 N=4である、請求項3に記載のハード・ディスク記憶システム。

【請求項5】 さらに、

(a) 前記各スピンドル駆動手段にตอบสนองして前記各ディスクの当該の回転位置間の位相差を指示するスピンドル速度誤差信号を発生する手段と、

(b) 前記誤差信号にตอบสนองして、前記ディスクの回転速度及び相対位置を電子的にロックし、それによって前記の4台のドライブの間にインタリーブされたデータに、読出し及び記録のために同時にアクセスできるようにするスピンドル駆動制御手段とを含む、請求項2に記載のハード・ディスク記憶システム。

【請求項6】 さらに、前記の4台のドライブの各々から対応するデータ区画を読み出し、前記区画を最初に記憶されたデータに変換するための制御手段を含む、請求項2に記載のハード・ディスク記憶システム。

【請求項7】 前記制御手段が、前記ドライブのうちの3台に記録されたデータ区画、及び前記ドライブのうちの第4のドライブに記録されたパリティ情報にตอบสนองして、前記ドライブから読み出された前記データに関するパリティ検査を実行するパリティ検査手段を含む、請求項6に記載のハード・ディスク記憶システム。

【請求項8】 さらに、前記ドライブのうちの任意の3台に含まれる対応するデータ・ビットにตอบสนองして、最初に記憶されたデータを再生するエラー訂正手段を含む、請求項6に記載のハード・ディスク記憶システム。

【請求項9】 さらに、

(a) 前記の4台のドライブのうちの任意の1台がハード・ディスク記憶システムから遮断されたことを検出し、前記ドライブのうちのどれが遮断されたかを指示する遮断信号を発生する遮断信号手段と、

(b) 前記の4台のドライブのうちの任意の1台がハード・ディスク記憶システムに再接続されたことを検出し、前記ドライブのうちのどれが再接続されたかを指示する再接続信号を発生する接続信号手段と、

(c) 接続されたままであった前記ドライブのうちの3台から対応するデータ区画を読み出し、前記の遮断されたドライブ上に最初にあったデータ部分を再生し、前記の再生された部分を前記の再接続されたドライブ上に記録する再生制御手段とを含む、請求項4に記載のハード・ディスク記憶システム。

【請求項10】 さらに、前記の遮断信号にตอบสนองしてハード・ディスク記憶システムを動作不能にする手段を含む、請求項9に記載の前記ハード・ディスク記憶システム。

【請求項11】 さらに、前記の再接続信号にตอบสนองして前記の動作不能にされたシステムを動作可能にする手段を含む、請求項10に記載のハード・ディスク記憶システム。

## 【発明の詳細な説明】

【0001】

【産業上の利用分野】 ディスク記憶装置は、コンピュータで使用する大量のデータ及びプログラムの情報を記憶する問題に対する好ましい手法である。ディスク記憶装置は、30年以上前に登場して以来、より大きな記憶容量、より短いアクチュエータ・アクセス時間、より短い回転待ち時間、及び信頼性を高めるためのディスク故障からの回復を提供するために継続的に改良されてきた。これは、このような装置の占める空間及び消費電力を減らしながら達成されてきた。

【0002】

【従来の技術】 ディスク記憶装置の性能の向上は、データ処理システムの他の面の改良に追いつけなかった。最近では、アクセス時間は1/2または1/3に低下し、データ伝送速度は2倍に向上したが、データ処理装置とディスク記憶システムの性能のギャップはますます拡大してきている。これは、大部分は、機械的構成部分をより高速度で移動させるにはより大きなエネルギー量が必要であるという物理的法則の存在によるものである。

【0003】 ディスク記憶システムの全体的性能は、典型的なデータ量、たとえば16キロバイト (KB) のデータを取り出すのに必要な時間の量によって特徴づけることができる。最近5年間で、この時間は1/3に低下したが、同じ期間に、システムの半導体部品の性能は10倍に向上した。この不均衡は、改善する兆しがなく、悪化し続けそうである。

50 【0004】 多数の端末にサービスするために通常使用

される中規模データ処理システムに関して性能を見ると、時間の約60%は記憶サブシステムによって消費されている。通常、1回のユーザ・トランザクションは、約20回のディスク動作を必要とし、各ディスク動作で約4KBのデータが転送される。このタイプのシステムは、端末操作員への応答時間が1-2秒である。この応答時間は、通常にみられる実際の値であるが、通常の端末操作員にとってこの値は理想的ではない。

【0005】この応答時間は、記憶サブシステムの性能によって大きく制限され、半導体装置とディスク記憶システムの間の性能ギャップが拡大し続けるにつれてますますそうなる。将来のアプリケーションがますます複雑になり、より多くのデータの転送が必要となると思われるため、この問題の深刻さがさらに増す。ユーザ・トランザクション当たりのディスク動作回数は、20回から40回に上昇すると予想され、平均データ転送速度は4KBから20KBに増加すると思われる。

【0006】データ処理システムが日常のビジネス及び個人生活のより不可欠な部分になりつつあるため、この問題はさらに複雑になっている。したがって、ユーザが修理のために熟練したサービス担当者を1日程度待つことは、もはや便利でもなく、可能でもない。大抵の場合、システムは、数分あるいはせいぜい1-2時間のうちにサービスできるように復旧しなければならない。そうしなければ、貴重なビジネスの機会が失われる。これは、故障した装置の障害追求及び修理または交換がユーザ自身によって行なわれるようになることを意味する。

【0007】ディスク記憶装置の回転待ち時間、サイズ、電力消費量、データ伝送速度、ビット密度、トラック密度、保守容易性、及び信頼性という設計上の問題に対する各種の手法がこれまでに用いられてきた。米国特許第3876978号は、複数の装置をもつ記憶システムを記述している。複数の装置の1つは、パリティ・ビットを記憶するのに利用される。このパリティ・ビットは、対応するビット位置にある当該の各データ用の装置に基づくものである。上記特許の記述によれば、この記憶システムは、1つのカートリッジを使って、他のすべてのカートリッジ内の対応するビットの比較から生ずるパリティ・ビットを表すという、テープ・カートリッジ・システムに関する。一つのカートリッジが失われた場合にそのデータを再生するため、すべてのデータ・カートリッジからすべてのビットを読み出し、パリティ・カートリッジ内の対応するビットの結果を組み合わせることで比較し、適正なパリティを生ずるようにしてデータを再生することができる。上記特許で採用された手法は、アクセス機構が1個しかない場合であるが、多重ディスク・ドライブ・システムに使用するのに好適である。ディスク・ドライブ・システムでは、失われたデータを再生するとともに、記憶システムから読み出されたすべてのデータの妥当性のオンライン検査を実行することが望まし

い。上記特許に記述された記憶システムを使用しても、ディスク記憶システムから読み出されたデータのオンライン検査に関して満足な性能は得られない。

【0008】米国特許第4036659号は、ディスク・ドライブ・システム用のプログラム式制御装置を示している。複数のディスク・ドライブの間でデータを分散することについて長々と詳しく記述されているが、超高速の回転速度で動作する例外的に小さいディスクを使用することは示唆されていない。

10 【0009】米国特許第4577240号に記述されたビデオ記録システムは、別々のアクチュエータを使ってエラー訂正またはディスク・トラブルからの回復の問題を記録するディスク記憶システムに関する。その代わり、他の方法ではエラーを生ずるはずの欠陥を含むトラックは、単純にスキップされる。実際のデータ処理の場では、こうすると、データ容量が容認できないほど失われるはずである。というのは、単一ビット・エラーが1トラック全体の損失をもたらすからである。これは、データ記憶の問題に適用するとき、容認できないトレード

20 オフである。上記特許はまた、複数のドライブ・スピンドルの使用を示唆しておらず、そのかわりに別々のアクチュエータを使用している。この手法は、故障した装置の簡単な交換には役立たない。というのは、2つのアクチュエータ、したがって2つのデータ・グループが関与するからである。

【0010】1台のディスク記憶システムに複数のスピンドルを使用することは、米国特許第4583133号で示唆されている。しかし、このシステムでは、一度にただ1つのドライブが使用されるだけである。データをすべての装置の間に同時に配分し記録することができるの教示はない。ハード・ディスクではなく可換性媒体の使用が意図されており、回転速度が可換性媒体では通常の低速度ではないとの言及はない。

30 【0011】米国特許第4724495号の教示は、連続したビデオ・フィールドをディスク記憶スタックの異なるゾーン上に記録する、別々の2つのアクチュエータを使用したビデオ記録システムを対象とする。別々のスピンドルが使用され、データが複数のスピンドル上に同時に記録されるとの示唆はない。このシステムは、トラック内の欠陥を、そのトラック全体を単にスキップすることによって処理する。これは、上記の米国特許第4577240号に示された手法に類似している。上記特許は、複数の小型ディスクが、データがすべてのディスクに並列に分散されている1つのシステム内で組み合わせられることを示唆していない。

40 【0012】このような装置は魅力的な特徴をもつにもかかわらず、まだ克服されていないいくつかの本質的な制限がある。たとえば、データが円形トラックに沿って直列に配列されるため、所望のデータがデータ・トランスジューサの下を通過するまで待つ必要がある。単純な

統計的観点からみると、所望のデータがトランスジューサに到達するのに必要な平均時間は、ディスクが1/2回転するのに要する時間である。セクタを1つのトラックから別のトラックへスキューすることによって、この単純な統計的平均待ち時間をやや改善することが可能である。こうすると、トランスジューサは、トラック・アクセス後にこれ以外の方法の場合より多少早くトラックの読出しを始めることができる。データがトランスジューサの下にくるまでの待ち時間は、回転待ち時間と呼ばれる。

【0013】トラックからトラックへのアクセス中にヘッドアーム・アセンブリを機械的に移動させるのに必要な時間は、より高速のアクチュエータを使用して大幅に短縮されてきたが、回転待ち時間はほとんど改良されていない。これは、ディスク・ドライブの仕様を調べると容易に理解される。いわゆる「ハード」ドライブ、すなわち剛体基板を利用した駆動機構は、必ず3600rpmで回転する。回転待ち時間は、回転速度に反比例するので、回転速度を増加させない限り大幅な改良は不可能である。

【0014】スピンドル駆動モータをスピードアップし、それによって回転待ち時間を短縮することは、簡単なことのように思われるが、より高い回転速度のディスク・ドライブがないことから明らかなように、そう簡単ではない。明白な解決には役立たないが、問題のいくつかは明かである。たとえば、ヘッドの空気力学的制御を行なってヘッドとディスクの間隔を20-63、5mmの範囲に維持する、いわゆる「ウィンチェスタ」技術は、毎秒1.5-2.5mの速度で移動する空気の薄い膜の存在を利用している。3600rpmで回転する、直径が約89mm-203mm(3.5-8インチ)以上のディスクでは、この速度要件は容易に満たされ、空気力学的パッケージ内に配置されたトランスジューサである「スライダ」を、これらのパラメータを満たして動作するように設計することができる。

【0015】回転待ち時間は回転速度に直接的に関係し、回転速度を増加させると回転待ち時間の短縮で直接の利益が得られることは以前から認識されてきた。それにもかかわらず、ほぼすべてのハード・ディスク駆動システムの回転速度は、3000-3600rpmの範囲にとどまっている。

【0016】これは、たとえば回転待ち時間を相当に改善するために回転速度を3倍にして約10,000rpmにすると、既存のスライダ技術が全く使えなくなることによって少なくとも部分的に説明される。そのようにすると、増加したディスク速度から生ずる従来とは異なった空気力学的状況に対処するために、コストのかかるスライダの再設計が必要となる。

【0017】空気力学の問題は、予測可能であり、十分な工学的努力によっておそらく解決できるはずである

が、他の問題についてはそうはいかない。適切な事前措置が講じられていない場合、ヘッドとディスクの偶然的接触が、装置が寿命に達する前に、故障を生じ得ることが認識されてきた。ディスクの回転開始及び停止時に、トランスジューサを、データが保存されない「着陸」ゾーン上方で位置決めする機構を設計することは可能である。こうすると、データが記憶されるディスク表面の摩擦及び潜在的な損傷の一部がなくなる。それでもなお、通常のデータ読出し及び書き込み中にトランスジューサとディスクの不測の接触が起こる可能性がある。したがって、トランスジューサを着陸ゾーンに位置決めするための特別の機構のコストが経済的に引き合う場合でも、ディスク表面に何らかの形の保護機構を設けることが望ましい。通常、この保護は、ディスク上の記録媒体の表面に塗布される潤滑剤の形をとる。

【0018】適当な潤滑剤の開発は困難であった。必要な潤滑特性の点から、二三のクラスの潤滑剤を除いてすべての潤滑剤が除外される。さらに、潤滑剤は、通常のビジネス環境または家庭環境で見られる汚染物質と干渉すると、通常のヘッドとディスクの間の間隔の実質的部分に取り付く分子、結晶、またはアモルファス構造物を形成するが、このような物理的構造物を形成する汚染物質と相互作用してはならない。ヘッドがこのような構造物と接触すると、「ヘッド・クラッシュ」を起こして、ディスク表面の損傷及びデータの損失をもたらす可能性がある。潤滑剤は、潤滑作用及び化学的性質の点から満足のいくものでなければならぬだけでなく、ディスク表面に付着し、回転によって飛び散らない物理的特性ももたなければならない。望ましい化学的特性及び潤滑特性を有するいくつかの潤滑剤は、遠心力の結果ディスクの周辺へ移行する傾向があることが判明している。この障害は、長期間使用後に始めて出現するので、満足のいく性能をもつかどうか潤滑剤を評価し試験するのは困難である。3,600rpmの回転速度で満足なディスク寿命をもたらす働きをする潤滑剤であっても、より高い回転速度で許容できるかどうかを確認するために少なくとも長時間の寿命試験が必要となる。このような試験の結果、おそらく既存の潤滑剤はずっと高速の回転速度での使用には満足でないことが示されるはずである。

【0019】したがって、回転速度が3,600rpmに固定され続けたことは理解できる。というのは、より高い回転速度から生じる予見できるが予測不可能な問題が、開発投資の実質的な抑止力になっているからである。

【0020】より高い回転速度を試みた場合、ドライブ内で発散される熱が問題になることが予想される。より高い回転速度でディスクを回転させる場合、必要な電力量が増加するので、必然的にスピンドル駆動モータがより大型になり、それに伴う電力損が増大する。これは、電力損の増大がきわめて小さな容積内で生じ、熱の放散

が一層困難になることを認識すれば、小さな問題ではない。また風損が増大し、やはり装置内で発生する熱を増加させることが予想できる。

【0021】占有空間は常に考慮すべき問題なので、より大型のドライブ・モータはより高い回転速度の使用に対するかなりの抑止力になる。ディスク・ドライブがすでに使用可能な空間及び電力のかなりの部分を占めている、いわゆるパーソナル・コンピュータの場合には特にそうである。パーソナル・コンピュータがますます小型化する傾向のもとでは、既存のディスク・ドライブより大きな空間を占めるディスク・ドライブは受け入れられない。さらに、より高い回転速度の装置のヒート・シンクのための空間が余分に必要になると思われる。

【0022】たとえ、回転速度を増加させて回転待ち時間を改善することができたとしても、このような開発には必然的にアクチュエータ・アクセス時間を改善する努力が必要となり、やはり装置内で消費される電力を増加させる傾向にある。

【0023】従来技術は、回転待ち時間を改善する努力とディスク・ドライブに対する他の要件との間に矛盾が存在するという十分な証拠を示している。回転待ち時間を改善した装置の電力とサイズが増大することは、装置をより小型化し、電力消費量を減らし、動作温度をより低くするという要件と矛盾する。とりわけ、装置の信頼性を高め価格を下げるのが本質的な要件である。

【0024】従来技術ではうまく対処されなかった別の問題は、欠陥ディスク・ドライブの修理または交換に関するものである。この作業には、従来、平均的なパーソナル・コンピュータ・ユーザの熟練を超える高度の熟練が必要であった。ハード・ディスク・ドライブの修理は、記憶されるデータの性質の故に、精巧度のより低いフレキシブル・ディスク・ドライブよりもクリティカルである。通常、ハード・ディスクは、システムで使用されるすべてのアプリケーション・プログラムを格納し、大量のデータを含むこともある。ハード・ドライブ上の情報はクリティカルなので、ハード・ディスク上のデータをフレキシブル・ディスクまたはテープ型装置上に複製する、「バックアップ」動作を定期的に行うのが習慣となっている。ハード・ディスクが損傷を受けた場合、またはそれ以外の原因で動作不能になった場合に、フレキシブル・ディスクまたはテープを使ってデータを回復することができる。

【0025】ディスク・ドライブは、誤ったデータに関するエラー訂正が実行できるように設計されているが、このようなエラー訂正能力は、習慣的にきわめて制限されており、小さな規模のヘッド・クラッシュから生ずる大きなデータ・ブロックの損失は処理できない。

【0026】バックアップ動作は実行が面倒であり、バックアップが必要なことは、ハード・ディスク・ドライブが故障するまでしばしば無視される。故障したときは

手遅れで、データは回復できない。ディスク・ドライブの交換は、しばしば複雑な作業であり、ディスク・ドライブの取り外しと交換の機械的側面、ならびにディスクをフォーマットし、まだ使用できる情報を再ロードするためのソフトウェア知識において、かなりの専門知識が必要となる。

【0027】従来技術のディスク記憶システムに伴う別の問題は、ディスク基板に使用される材料に関するものである。アルミニウムが、この技術の出現以降ほぼずっと選択されてきた材料である。アルミニウムは、軽量、優れた加工性、及び強度という利点を提供する。以前の値の磁気媒体の厚さ及びヘッド飛行高さでは、アルミニウム基板は、費用はかかるが通常の機械加工技術を使用して適切な仕上げにすることができる。しかしながら、最先端の加工装置を使用しても、アルミニウム基板の表面仕上げは、ビット密度及びトラック密度を向上させるのに必要な薄い媒体コーティングに対する許容差を超える欠陥を含む。

【0028】エラーのないディスクを提供するには、表面仕上げが完全でなければならない。ディスクを磁気媒体でコートしてから、その表面をきわめて微細な表面に仕上げることができる。こうすると、ヘッド・クラッシュに伴う問題の深刻さは軽減されるが、基板内に微細な突起またはビットが存在すると、たとえ媒体の表面がほぼ完全であっても、突起領域内でのビット抜け、及びビット領域内でのビット拡散を生ずる。アルミニウム基板の表面を仕上げるための最先端の技法は、材料の理論的限界に近づいている。純粋なアルミニウムを使用すると、不純物、及びアルミニウム合金に良くみられる包有物を含まないで、良好な仕上げを行なえる可能性が高くなる。残念ながら、純粋アルミニウムはきわめて軟らかく、加工性には乏しいので、得られる表面の品質が制限される。基板に使用するために他の材料も評価した。ガラスと半導体用シリコンが使用されてきた。どちらの材料もアルミニウムよりもずっと良好な表面仕上げを行なえる可能性がある。セラミック基板も使用できる。しかしながら、これらの代替物は、アルミニウムよりも脆い。この特徴のために、これらの材料の広範囲の使用が妨げられてきた。素人による頻繁な移動に耐えられるように比較的頑丈にしなければならないパーソナル・コンピュータで使用されるタイプのドライブでは特にそうであった。さらに、セラミックス、ガラス、及び半導体シリコンの機械的強度は、現在の回転速度で動作の衝撃及び応力に耐えるには十分であるが、回転速度がずっと高くなったときには疑わしい。

【0029】トランスジューサを所望のトラックの上方で位置決めするアクチュエータの性能が絶えず改善されて、アクセス時間がより速くなってきたが、この性能の向上は主として電力消費量の増加という犠牲を払って達成された。これは、電源の容量の増大、バッテリー駆動シ

システムの動作時間の短縮、及びドライブ内で発生する熱の増加の点で、コンピュータ・システム全体にとって問題を生ずる。

【0030】高性能ディスク・ドライブを使用する大型システムの性能は、各アクチュエータからアクセスできる膨大な量のデータによって課せられる制約に達することが判明している。このような各ドライブ上のデータへのアクセスは必ずしも逐次的には進まないで、データの検査に伴う遅延が避けられない。この制限は、各アクチュエータの下に置くデータを減らすことによって避けることができるが、各アクチュエータ及びそのディスク・ドライブ・システムの関連部品が高価なので、このような手法は経済的に実用的でない。言いかえると、高価なアクチュエータは大量のデータと組み合わせなければ、それを利用する意味がない。高速アクチュエータで比較的少量のデータにアクセスするという目標は、これまで商業的環境では実現不可能であった。

【0031】さらに、各アクチュエータの下に置くデータを減らすことの利点は認識されてきたが、この問題に対する従来の手法は、追加のアクチュエータの存在により信頼性が失われるという問題があった。

【0032】要約すると、既存の装置より回転待ち時間を改善し、より小型で、より小さな電力を使用するディスク・ドライブ・システムを実現することが望まれる。この装置は、既存ディスク・ドライブ技術をできるだけ多く利用して既存の潤滑剤及びスライダ技術を使用できるようにするのが理想である。これらの改良は、アクチュエータ性能の向上を伴い、その結果回転待ち時間の短縮を伴うのが理想である。これらのことをすべて、信頼性を低下させずに達成しなければならない。このシステムは、顧客が、複雑なツール、平均的なユーザの能力を越える特別の機械的熟練またはプログラミング能力を使用せずに、ディスク記憶システムの欠陥部分を交換できるようにすることが好ましい。

【0033】

【発明が解決しようとする課題】本発明の目的は、単一ドライブ・システムよりも信頼性の高い多重ドライブ式ディスク記憶システムを提供することである。

【0034】本発明の別の目的は、多重ドライブ式ディスク記憶システムを電力遮断する必要なく、かつデータの損失なく、個々のドライブを接続及び切断できる、多重ドライブ式ディスク記憶システムを提供することである。

【0035】本発明の別の目的は、4つのドライブをもち、そのうちの1つのドライブが故障してもデータの損失を起こさない、または実質的に性能を損なわない、ディスク記憶システムを提供することである。

【0036】本発明の別の目的は、データの並列検索によってデータがより高速で検索できるように、高速、低コストのアクチュエータを使用して比較的より少量のデ

ータにアクセスすることのできる、ディスク記憶システムを提供することである。

【0037】本発明の別の目的は、アクチュエータ速度の損失またはシステム信頼性の低下を起こさずに、各アクチュエータによってアクセスできるデータの量を減らす、ディスク記憶システムを提供することである。

【0038】

【課題を解決するための手段】N個のドライブのうち、第1ないし第N-1番目のドライブにデータを記憶させ、第N番目のドライブにパリティ情報を記憶させたハード・ディスク記憶システムが提供されるので、1つのドライブに故障があったり遮断されたりしてもデータが失われない。また、例えば、標準的な113mm(5.25インチ)のディスク・ドライブに通常必要な空間内にN台、例えば4台のユニットを装着することができる。4台のユニットが使用できるので、各ドライブ上に、情報の各バイトの一部分、ならびに4台のドライブのうち任意の3台からデータが再生できるのに十分なエラー訂正情報を割り当て、記憶することができる。こうして、各アクチュエータの下にデータ量がより少ないという利点が、通常なら複数ドライブの使用から生ずるはずのトラブルの頻発なしに、保持される。

【0039】

【実施例】本発明は、従来技術と著しく異なる機能構成を含む。図1において、ディスク記憶システム1の機能構成は、アレイ／クラスタ制御装置5、及び複数の記憶機構アレイ7、8、9を含む。前記の各アレイは、4つの分離したディスク・ファイル、たとえば7a、7b、7c、7dを含むことが好ましい。各アレイ内の別々のディスク・ファイルの実数の数は、4より大きくてもよいが、前述したように半高ディスク・ドライブに対する既存のパーソナル・コンピュータ形状因子との幾何的整合性の理由から4が好ましい数である。本発明の構成では、アレイは、直接アクセス記憶装置(DASD)サブシステム設計である。この設計は、追加の機械的部品をもつより複雑な設計から通常予想されるものとは違って、データの記憶と取出しがより高速かつ高い信頼性で行なわれるような方式で、データをシステムの構成要素の間に分散する。

【0040】データは、データ・バス6によって中央演算処理装置3に接続されているアレイ／クラスタ制御装置5の監督下で、ディスク記憶サブシステム1とデータ処理システム2の間で受け渡される。制御装置5は、データ処理システム10から受け取って、データ処理システム2のランダム・アクセス・メモリ4に記憶したデータの形式を、アレイにデータを記憶する際に使用する形式に変換する働きをする。制御装置5は、データのパーセルを3つの単位+パリティ単位にセグメント化し、制御線20上の信号によって選択されたアレイのドライブa、b、c、d上に記憶するために、得られた4つの



単位を、データ線21a、21b、21c、21dを介して転送する働きをする。

【0041】図2は、データ形式が、データ処理システム2内で使用される形式から、本発明のディスク記憶システムの形式に変換される様式を示す。データ処理システム2のランダム・アクセス・メモリ4に記憶されたデータは、通常、各522バイトのパーセル圧縮データ・ブロック100、101ないし10nに配列される。アレイ/クラスタ制御装置5は、図2に示したように動作して、各データ・ブロックを各174バイトの3つの区画100-1、100-2、100-3に分割する。区画100-1、100-2、100-3の各々の第1バイトは、バッファ101-1、101-2、101-3のうち対応する1つに置かれる。これらのバッファから、対応するデータ単位は、逐次、線103-1、103-2、103-3を介して対応するバッファ及びエラー訂正コード(ECC)生成機構105-1、105-2、105-3に同時に供給される。データ・バッファ101-1、101-2、101-3からの3つの単位は、線104-1、104-2、104-3を介してパリティ生成機構ブロック110に供給され、バッファ及びECC生成機構105-4につながる線103-4上に適当なパリティ単位を発生する。

【0042】線106-1、106-2、106-3、106-4上で得られる4つの単位は、データ・ブロックから最初に読み出された3つの対応するデータ・ブロック部分+1つのパリティ・ビットを表す。これら4つの単位は、4つのDASD装置100a、100b、100c、100dの同じ論理アドレスに記録される。各単位は、異なるディスク・ドライブに記録されるが、並列(同時)に記録される。データ・ブロック100の転送が完了すると、そのデータ・ブロックの第1の単位100-1はDASD100-aに記録され、第2の単位100-2はDASD100-bに記録され、第3の単位100-3はDASD100-cに記録され、データ・ブロック100のパリティ単位はDASD100-dに記録される。

【0043】記録されるデータは、通常、522バイト・ブロックで制御装置5に転送される。522バイトのデータ・ブロックの3つの174バイト・データ区画への変換は、そのデータ・ブロックの各区画を論理FIFOバッファに読み込むことによって実行することが好ましい。このFIFOバッファは、バッファ101-1、101-2、101-3でよい。

【0044】欠陥のあるまたは動作しないDASD上で失われたデータを再生する技法が、図3に示されている。この図では、データ部分100-3を含むDASD100cが、何らかの形で故障して、それに含まれるデータが利用不可能になったと仮定する。これは、図3で、バッファ及びECC生成機構105-3からデータ

・バッファ101-3につながる線103-3上のXによって図示されている。DASDの故障を検出するための手段については、後で論じる。DASD脱出し障害が検出されると、システムは図3の配置を取るように再構成される。図3の配置は、図2のそれと本質的に同じであるが、データの流れの方向が逆である。各単位がDASD100a、100b、及び100dから利用可能になると、バッファ及びECC生成機構105-1、105-2、105-4によって、元のデータ形式に変換される。DASD100cからの単位は入手できない。線103-1、103-2、103-4上の信号が、パリティ・生成機構110に供給され、そこで失われた単位が再構築されて、線103-3aを介してデータ・バッファ101-3に供給される。線103-3aは、DASD100cが故障したことをシステムが認識した後で、そのシステムによって行なわれた接続を表す。破線103-2a及び103-1aは、それぞれDASD100bまたは100cが故障した場合に行なわれる接続を表す。

【0045】好ましい実施例では、データ区画、パリティ単位、及びエラー訂正コードを発生するタスク専用のハードウェアとソフトウェアの組み合わせを利用するが、これらの同じ機能を、適当なプログラムの制御下で動作するマイクロプロセッサ、または完全に専用のハードウェアによって実行することもできる。ほとんどの場合、最適な構成は、専用ハードウェアならびにソフトウェアを含むものである。具体的な実施例は、データ転送速度とマイクロプロセッサの速度の関係、代替各手法の相対的成本、及びその他の設計上のトレードオフなどの因子に応じて変わる。

【0046】本発明の好ましい実施例では、522バイト・ブロックを3つの174単位区画に変換することを企図していることを理解されたい。これらの区画は、次に4つの別々のディスク・ドライブにパリティ単位とともに並列に記録される。これが最適な配置であると考えられるが、他の構成を使用してもよい。たとえば522バイト・ブロックを3つの区画に変換する代わりに、各バイトを3つの部分に変換し、または522バイトより大きいまたは小さいブロックを使用することも可能である。これらの、またはその他の修正を加える場合、バッファにも修正を加えるのが適切である。最適サイズから外れると、バッファとディスク・ドライブの間のデータの転送が最適より悪くなる結果、性能が低下することがある。他の構成としては、4つより多い、または少ない数のディスク・ドライブを使用することがある。ただし、4という最適のドライブ数から外れると、半高フレキシブル・ディスク・ドライブに対する工業標準との物理的整合性の利点が失われて、最適な4ドライブ構成よりも信頼性が低下することがある。

【0047】図1に戻ると、アレイ/クラスタ制御装置

5は、システム内の各DASDに関する状況情報を維持するという追加の機能をもつ。この機能は、プログラム制御下で動作するマイクロプロセッサによって通常の方法で実行することが好ましい。状況情報は、「DASDリスト」、「性能低下DASDリスト」、及び「DASD再構築リスト」という3つのデータ・セット内に維持される。DASDリストは、システム内のすべてのDASDに関するすべてのバイタル(重要)プロダクト・データを含み、クラスタ内のDASDの位置、すなわちM行×N列に編成される。バイタル・プロダクト・データとは、元々DASD内に含まれ、アレイ/クラスタ制御装置5によってそのDASDから取り出される情報である。このような情報は、通常、そのDASDの製造連番号、技術変更の数または技術変更レベルあるいはその両方を含む。この情報は、電力投入時及びその他の非常時に、システム内のすべてのDASDをポーリングすることにより制御装置によって読み出される。この手順によって、故障して交換されたドライブに関するバイタル・プロダクト・データが、新しいDASDを反映するように確実に改訂される。新しいドライブの検出は、単に、各ドライブから読み出されたバイタル・プロダクト・データを、システム内に以前に存在した(DASDの最後のポーリングによって発生された)バイタル・プロダクト・データと比較することによって実行できる。あるDASDがアレイ/クラスタ制御装置5から出されたポーリングに回答しない場合、そのDASDに、破壊または動作不能のフラグが付けられる。変更されたことが検出されたドライブは、そのドライブに割り振られたデータが再構築されるようにフラグ付けされる。故障したDASD上に以前に含まれていたデータを再構築するための手順については、先に言及したが、後でより詳細に説明する。

【0048】性能低下DASDリストは、DASDが故障したが、故障した装置を直ちに交換したくないときに使用する。すでに述べたように、DASDが故障しても、システムが完全に動作不能になるのではない。すなわち、故障したDASDに含まれていたデータは、図3に関して説明したように再生することができる。故障したDASDに含まれていたデータの再生は、欠陥装置の代りに使用される新しいDASD上にそのデータを再書き込みするために行なうこともでき、また再生されたデータを、あたかも故障したDASDから直接来たかのように、データ処理システム2が使用することもできる。エラー回復ルーチン後にあるDASDがバイタル・プロダクト・データ読出しコマンドまたは書き込みコマンドを完了しない場合、そのアレイ内のその特定のDASDに性能低下のフラグが付けられる。データ処理システムは、性能が低下した装置に読出しコマンド及び書き込みコマンドを出したとき、性能低下DASDリストを利用して、通常エラー回復ルーチンを阻止することができる。

このような状況では、図3のデータ回復技法によって、データ処理システムは、転送されたすべてのデータについてエラー回復機能を実行するという負担を課されずに、引き続き機能することができる。もちろん、この能力に対して対価を支払わなければならない。すなわち、同じアレイ内の別のDASDが故障すると、そのシステムの動作が継続できないという状況が生ずる。故障したDASDをもつシステムの動作は限られた時間しか続けず、故障したDASDをできるだけ早く交換することが望ましい。

【0049】アレイ/クラスタ制御装置5は、故障したDASDが動作可能な装置と交換されたことをバイタル・プロダクト・データから認識すると、故障したDASDを性能低下DASDリストから削除し、新しいDASDをDASD再構築リスト上に置く。このDASD再構築リストは、そのアレイ内の以前に故障した位置に割り当てられていたデータが再生されることを必要とする個々のDASDの記録である。DASD再構築リストは、故障したDASDの識別に加えて、データが復元されたDASD内のセクタの数をも含んでいる。セクタ情報を含むリストの使用によって、システムは、データをバックグラウンド・モードで再生することができ、それによって、中央演算処理装置3による通常コマンドの実行が可能になり、データ処理システム2の動作に対する影響を最小にすることができる。

【0050】アレイ/クラスタ制御プログラムの、図4のデータ流れ図で表される部分では、サブルーチンは、ブロック40から開始し、システムが電力投入されたことを表す信号、またはシステムがリセットされたことを表す信号を検出する。これらの信号は、通常、データ処理システム内で発生され、さまざまな目的に使用される。

【0051】ブロック41で、クラスタ内のすべてのDASDをリセットし、そのクラスタ内の各アレイ内の個々のDASDをポーリングする。このポーリングによって、そのクラスタ内の各動作可能DASDからバイタル・プロダクト情報が読み出される。各DASDは、その独自の連番号、及び技術変更または技術変更レベルあるいはその両方を示す情報を供給する。さらに、そのDASDの個々のセクタに関する情報も提供される。

【0052】クラスタの各アレイ内のすべての位置のポーリングが完了すると、ブロック42でテストを実行して、そのクラスタ内のいずれかのDASD位置がそのポーリングに回答しなかったかどうか判定する。あるDASD位置がポーリングに回答しないと、エラー状態と解釈され、プログラムはブロック42aで再試行/エラー回復手順に分岐する。この手順は、ブロック41で表されるポーリング中に適切に回答しなかったDASDからバイタル・プロダクト・データを読み出そうと試みる。

ブロック42aの再試行/エラー回復段階で、前に故障

していたDASDからバイタル・プロダクト・データを読み出すのに成功した場合、ブロック43bでNO分岐を取って主命令シーケンスに戻る。一方、適切に回答しないDASDがまだある場合は、ブロック43bでYES経路を取ってプログラムはブロック42cに分岐することができる。ブロック42cで、回答しなかったDASDを、それがまだ削除されていない場合、DASD再構築リストから削除する。さらに、故障したDASDを、性能低下DASDリストに追加し、その後、主プログラムに戻る。

【0053】利用可能なすべての情報が動作可能DASDから読み出され、故障したDASDが性能低下DASDリストに追加されると、ブロック44で、プログラムは現バイタル・プロダクト・データを最後の読出しから得られたデータと比較する。DASDからのバイタル・プロダクト・データは、アレイ内のDASD位置に従って配列される。好ましい実施例では、各アレイ内に4つのDASD、記憶サブシステム内にN個のアレイがある。最後の読出し以降に情報が変更された場合、プログラムはYES経路を取ってブロック45aに分岐し、バイタル・プロダクト・データの変更を反映するようにDASDリストを更新する。次にプログラムは、ブロック45bで、故障したDASDをDASD再構築リストに追加する。さらに、ブロック45cで、再構築リスト上にあるすべてのDASDを性能低下リストから削除することによって、記憶されたDASD情報を更新する。

【0054】この時点で、ブロック46で示されるようにサブルーチンは完了し、プログラムはブロック47で、サブルーチンが完了し、CPUが記憶サブシステムにデータを転送し記憶サブシステムからデータを要求することができることを示す、「クラスタ・レディー」信号を発生する。次にサブルーチンは、ブロック48で、次の電力投入またはシステム・リセット信号が検出されるまで、遊休状態に入る。

【0055】第5図には、システムの電源を遮断しないで欠陥DASDの交換を行なうための技法が示されている。欠陥DASDを交換するためにシステムの電源を遮断する方が簡単であるが、それが望ましくない場合が多数ある。たとえば、システムは、欠陥DASDがあっても、限られた動作をすることができることがあり、電源を遮断すると、この限られた動作が完全に停止することになる。これは、システムがリアル・タイム動作に関係している適用業務、たとえば制御端末やデータ収集システムではきわめて望ましくない。このような環境では、たとえ正常のDASD性能が得られなくても限られた時間動作を継続することがしばしば可能であり、システムを停止するときわめて深刻な結果が生じる。したがって、たとえ限られた時間低い性能で動作することになっても、システムを停止せずに、システムの故障した構成要素の修理または交換が可能な技法が強く求められてい

る。

【0056】あるDASDが故障したとき、通常の処置は、できるだけ早く欠陥装置を交換することである。交換は、機械的にもプログラミングの点からもきわめて簡単に実行できるので、コンピュータ保守の熟練者の助けを借りずに、操作員の手でその操作を実行することができる。したがって、訪問サービスを要請し、保守担当員がコンピュータ設置場所に出かけてくるのを待つ遅延がない。こうすると、システムを完全能力に復旧する時間が節約されるだけでなく、訪問サービスの経費も節約される。操作員がシステムの電源を遮断し、欠陥ドライブの物理的交換を実行することも可能かもしれないが、新しいドライブをオンラインにすることも必要である。それには、通常のパーソナル・コンピュータ・ユーザの熟練レベルでは容易に生えられないプログラミング能力が必要であった。

【0057】したがって、問題は3つある。欠陥DASDの交換が物理的に簡単でなければならず、DASDを交換した後でシステムを動作状態に復旧するのにプログラミングの熟練を必要としてはならず、システムは性能が低下しても欠陥ドライブで動作を継続できなければならない。

【0058】システムから電源を遮断しないでDASDを交換することは「ホット・プラグング」と呼ばれる。DASDがプラグ接続される各アセンブリは、プラグの一部分としてインタロック・ワイヤを含む。DASDをシステムから外すとき、ワイヤが接続されている回路は切断される。システムが外されたDASDからデータを読み出そうと試みる場合、すでに説明したように、バリエーション回路が失われたデータを再構築する。システムが外されたDASDにデータを書き込もうと試みる場合、システムはそのDASDに性能低下のフラグを付ける。

【0059】操作員がシステムをオープンして、欠陥DASDを外すと、インタロック回路がオープンされる。動作可能DASDをプラグ接続すると、インタロック回路が閉じて、割込み信号が発生して制御装置5に送られる。割込み検出は、データ流れ図のブロック50で行われる。次にプログラムは、ブロック51に進み、特定のDASDクラスタ/アレイ位置に関するバイタル・プロダクト・データ(VPD)を読み出す。ブロック52で、プログラムは、割込みが発生した位置に関するバイタル・プロダクト・データを調べ、それが性能低下DASDリスト上にあったかどうか判定する。DASD位置が性能低下リスト上にあった場合、プログラムはYES分岐を取ってブロック52aに進み、そのDASDを性能低下DASDリストから削除し、それをDASD再構築リストに追加し、ブロック54で主プログラムに戻る。

【0060】ブロック52で実行されたテストでNOの結果となった場合は、そのDASD位置が性能低下DA

SDリスト上になかったことを示し、ブロック53に分岐する。ブロック53で、プログラムは、新しくプラグ接続されたDASDからバイタル・プロダクト・データを読み出し、読み出されたデータをそのDASD位置に対して以前に存在したバイタル・プロダクト・データと比較する。バイタル・プロダクト・データが変更されている場合は、新しいDASDがその位置にプラグ接続されたことを示し、YES分岐を取ってブロック53aに進み、そのDASDをDASD再構築リストに追加し、その情報をDASDリストに追加し、ブロック54で主プログラムに戻る。あるいは、ブロック53のテストでNOの結果となった場合は、バイタル・プロダクト・データが変更されていないことを示し、プログラムは単にブロック54で主プログラムに戻る。

【0061】元のデータを新しいDASD上で再構築することは、図6のデータ流れ図で表されるプログラムによって実行される。このプログラムの入口は、ブロック60で表されるアレイ/クラスタ制御装置5プログラム内の遊休ループからである。ブロック61で、プログラムはDASD再構築リストに照会して、データを復元しなければならないDASDがあるかどうかを調べる。このテストは、周期的に、またはアレイ/クラスタ制御装置5が遊休状態にある期間だけ実行することができる。ブロック61のテストで再構築を必要とするDASDがあることが示された場合、YES経路から分岐する。この場合プログラムはブロック62に移り、DASD再構築リストを読み出す。もちろん、ブロック61のテストでNOの結果になった場合は、プログラムは、ブロック63で、単に遊休状態に戻る。

【0062】ブロック64で、プログラムは、コマンド割込みを禁止する。これによつては、セクタのブロックの再構築が読出し/書き込み動作によって影響を受けないようになる。これは、最も簡単な再構築の方法であるが、他の手法も可能である。たとえば、アレイの残りの部分への通常アクセスを可能にしながら、再構築されるセクタだけをロードする方法がある。このような割込みは、再構築プログラムから出た後に再び許可される。

【0063】次にブロック65で、プログラムは、DASD再構築リスト上の各DASDに対して所定の数のセクタからなるブロックを再構築する。

【0064】ラウンド・ロビン方式でデータ、すなわち一度にx個のセクタの再構築を実行することによって、動作に必要なブッケーピングは大幅に単純化される。さらに、この構築は、システムの全体的動作に最小の影響しか与えないように実行するのが理想的である。主制御プログラムが制御を再確認できる前に1つのDASD全体が再構築される場合、必要な時間の量がシステムの性能に悪影響を与えることがある。DASD再構築プログラムへの1回のエントリによって再構築できるセクタの数の限度をもうけることによって、システムがこの作

業にさく時間の量も同様に制限される。

【0065】ブロック66のテストで、再構築の完了したDASDがあるかどうか判定する。これは、再構築されたセクタの数を調べることによって判定される。すべてのセクタが再構築されたDASDがある場合、YES分岐を取ってブロック66aに進み、DASD再構築リストからそのDASDを削除し、ブロック67で主プログラムに戻る。あるいは、ブロック66のテストでNOの結果となった場合は、プログラムはブロック67に直接進む。ブロック67で、DASD再構築リストに含まれる再構築されるセクタの数を更新する。次にブロック68で、プログラムは遊休状態に戻る。この状態からブロック61のテストが再び実行できる。

【0066】制御プログラムの、DASDアレイからデータを読み出すための部分が、図7に示されている。ブロック70でサブルーチンに入って、データ処理システム2から読出しコマンドを発行する。ブロック71で、プログラムは、選択されたアレイ内のすべてのDASDにトラック・シーク・コマンドを出す。ブロック72で、プログラムは、選択されたアレイ内のすべてのドライブ上の所望のデータ・セクタの行を読み出す。ブロック73のテストで、エラー訂正コードの検査から、またはコマンド・エラーとして、エラー指示があるかどうか判定する。

【0067】ブロック73のテストでNOの結果になった場合、データ中にエラーがないことを指示し、プログラムはブロック74に進み、アレイ内のDASDから読み出されたデータを組み合わせて、記憶サブシステム1に記憶するためデータ処理システム2から最初に渡された形式でそのデータを置く。

【0068】しかし、ブロック73でYESの結果となった場合は、プログラムはブロック73aに分岐し、そこでプログラムはエラー信号を分析して、単一セクタ・エラーが起こったかどうか判定する。そのエラーが単一セクタに限定された場合、YES経路を取って、プログラムはブロック73bに分岐し、そこでエラーのために失われたデータをセクタ・パリティ情報を使って回復する。次にプログラムは、主プログラムに戻り、DASDから取り出されたデータを組み合わせて、最初に記憶されていたデータ・ブロックを再生する。

【0069】ブロック73aからNO経路を取って分岐する場合は、エラーが2つ以上のセクタに広がっていることを示し、プログラムはブロック73cに分岐する。ブロック73cで、システムは読出し動作を再試行し、さらにデータ回復動作を実行する。これによってうまくエラーの大きさを単一セクタに縮小することができた場合、プログラムはYES経路を取ってブロック73dからブロック73bに移る。エラーの大きさが単一セクタより大きいままであった場合は、ブロック73dからNO経路を経てサブルーチンから出て、ブロック73eで

19

読出し中にハード・エラーがあったと報告する。

【0070】プログラムが、ブロック74への経路を取って元のデータを再生する場合、プログラムは続いてブロック75に進む。ブロック75で、読出しコマンドをテストして、読み出さなければならないセクタがまだ残っているかどうか調べる。読出し動作がまだ完了していない場合、YES経路を取って、プログラムはブロック72で読出しプログラムに再び入る。読み出さなければならないセクタがもうなくなった場合は、NO経路を取って、プログラムはブロック76に分岐し、読出し動作が首尾よく完了したことを指示する。

【0071】図8及び図9のデータ流れ図は、システムによって書き込み動作がどのように実行されるかを示す。データ処理システム2の中央演算処理装置3が、書き込みコマンドをアレイ/クラスタ制御装置5に出す。この書き込みコマンドは、メモリ・アドレス、記憶サブシステムに書き込まれるファイルの名前、及びデータが書き込まれるアレイを指定する。

【0072】ブロック80でアレイ/クラスタ制御装置5がデータ処理システムから書き込みコマンドを受け取ると、プログラムは、ブロック81に進んで、書き込みコマンドによって指定されたアレイを、性能低下DASDリスト上の性能低下DASDのリストと比較する。書き込みコマンドによって指定されたアレイが性能低下DASDリスト上にない場合、NO分岐を取ってブロック81aに進み、指定されたDASDアレイにトラック・シーク・コマンドを出す。指定されたアレイがこのコマンドを受け取り、アクチュエータが読出し/書き込みヘッドをディスク上の希望するトラックの上方に移動させる。シーク・コマンドを出した後、ブロック81bで、図2に示されたデータ・ブロック100などのデータ・ブロックを取り、それを適当な数、たとえば選択された例では3個の単位(区画)に分割し、その区画に対する新しいパリティ単位を生成する。ブロック81bで実行される動作は、先に図2に関して示し説明した、データ形式の交換及び関連するバッファ記憶に対応する。この動作は、システムで使用されるエラー訂正コードに従ったデータの変更も含んでいた。

【0073】次にブロック81cで、4つの並列経路を介して選択されたDASDアレイ内の4つのドライブに、変換されたデータを書き込む。これは、図2の信号線106-1ないし106-4によって示される動作である。データの書き込みに続いて、ブロック81dで、記録されたデータを検査して、書き込みエラーがあったかどうか判定する。これは、DASDから戻された状況コードを調べることによって実行される。トラック・エラーは、書き込み動作の前に検出される。エラーが検出された場合、プログラムは、YES経路を取ってブロック81eに分岐し、検出されたエラーの性質に対応するエラー回復手順を実行する。ブロック81fで、書き込みエラー

20

またはシーク・エラーがあるかどうか第2のテストを実行する。ここでエラーが検出された場合は、プログラムは、YES経路を取ってブロック81gに進む。ブロック81gで、ブロック81fで検出されたエラーの追加分析を実行して、2つ以上のDASDがエラーを発生したかどうか判定する。YES分岐を取る場合は、2つ以上のDASDがエラーを起こしていることを示し、YES分岐を取ってブロック81hに進み、ハード書き込みエラー、すなわち同じ位置に再書き込みをすることによって訂正されないエラーがあることを指示し、書き込み動作が試みられたDASDアレイを、性能低下DASDリストに追加する。

【0074】ブロック81gのテストで単一のDASDだけがエラー状態にあることが判明した場合、NO経路を取ってブロック81iに分岐し、故障したDASDを性能低下DASDリストに追加する。次にプログラムは、ブロック81jに進み、書き込みコマンドと、書き込まれたセクタの数を比較して、書き込むべきセクタがまだあるかどうか調べる。セクタがすべて書き込まれている場合、書き込みコマンドは完了し、プログラムはNO経路を取ってブロック82aに進む。ブロック82aで、書き込みコマンドが首尾よく完了し、次のコマンドの実行のために記憶サブシステムが使用できることを指示する。

【0075】ブロック81dに戻って、エラーがあるかどうかのテストでトラック・シーク・エラーまたは書き込みエラーが見つからなかった場合、プログラムはNO経路を取ってブロック81kに進み、書き込みコマンド中に存在していたセクタの数と、書き込まれたセクタの数を比較し、書き込むべきセクタがまだあるかどうか判定する。ブロック81kからYES分岐を取る場合、プログラムは、ブロック81bに戻って、次のデータ・ブロックを取りそれを分割してデータ単位及びパリティ単位を形成し、得られたデータを指定されたDASDアレイに書き込む。すべてのセクタが書き込まれている場合、ブロック81kのテストはNO分岐を取ってブロック81lに進む。このブロックは、ブロック82aと類似のものであり、書き込みコマンドが首尾よく完了し、次のコマンドの実行のために記憶サブシステムが使用できることを指示する。

【0076】次にブロック81に戻って、性能低下DASDリストを書き込みコマンドによって指定されたアレイと比較する。指定されたDASDがそのリスト上にあった場合、プログラムはYES経路を取って、ブロック83aに分岐し、プログラムの、データを性能低下モードで書き込む部分に進む。このモードでは、故障ドライブが性能低下DASDリスト上にあることは単に無視される、すなわち故障DASDは、あたかも正しく機能しているかのように取り扱われる。もちろん、これは、通常のエラー回復技法で故障したDASDが十分に補償でき

21

ると仮定したものである。このことについては、後でさらに説明する。

【0077】ブロック83で、先に説明したブロック81bと同様に522バイトの元のデータ・ブロックを各174バイトの3つの区画に分割する。プログラムは、ブロック83cに進み、先に説明したブロック81cと同様に選択されたDASDアレイにデータを書き込む。ブロック83cの書き込み動作は、あたかも欠陥DASDが実際に適切に機能しているかのように行なわれ、それが正しく機能していないことを無視する。

【0078】ブロック83dでテストを実行し、書き込みエラーまたはトラック・シーク・エラーがあったかどうか判定する。このテストでNOの結果になった場合、プログラムはブロック83kでテストを実行し、まだ書き込まれずに残っているセクタがあるかどうか判定する。ブロック83dに戻って、書き込みエラーまたはトラック・シーク・エラーが起こった場合、プログラムはブロック83eに分岐し、先に説明したブロック81eと同様にエラー回復手順を実行する。

【0079】次にブロック83fで、エラー回復手順の結果を検査して、未訂正のエラーがあるかどうか調べる。このブロックは、ブロック81fと同じ手順に従う。すべてのエラーが訂正されている場合、プログラムは、NO分岐を取ってブロック83kに進み、書き込みコマンドによって指定されたセクタを書き込まれたセクタの数と比較する。書き込むべきセクタがまだある場合、プログラムは、YES分岐を取って、ブロック83bに戻る。すべてのセクタが書き込まれている場合、プログラムはNO分岐を取ってブロック82aに進む。ブロック82aで、書き込みコマンドが完了し、別のコマンドを実行するために記憶サブシステムが使用できることを指示する。

【0080】ブロック83fのエラー回復手順でブロック83dで検出されたエラーから回復できなかった場合、プログラムはYES分岐を取ってブロック83gに進む。ブロック83gで、ハード書き込みエラー、すなわち回復不能なエラーがあることを指示し、そのエラーを発生したDASDを性能低下DASDリストに追加する。次にプログラムは、ブロック81で主プログラムに戻る。

【0081】図10は、個々のDASD装置のスピンドル・モータを制御するためのシステムを示す。すでに説明したように、DASDは、4つのDASDのアレイに配置されている。これらのアレイは、任意の数のアレイからなるクラスタに組み合わせられる。好ましい実施例では、各アレイは4台のDASDから構成される。区分されたデータは、一度に4単位ビットずつアレイに書き込まれる。すなわち、各区画から1単位ずつとそれに対応するパリティ単位（ブロック）が、同時に4枚のディスクに書き込まれる。読出しも並列に行なわれるので、ア

22

レイの4枚のディスクすべての回転位置を、互いに正確な関係に保たなければならない。そうしなかった場合、システム性能は低下する。すなわち、読出し動作は、最後のDASDがデータをバスしてしまうまで待たなければならない。したがって、4台のDASDが同じ速度で回転し、さらに互いに所定の位置関係を獲得し、記憶サブシステムの動作中ずっとこの関係を維持することが極めて望ましい。

【0082】この作業の困難さは、1トラック上に記録されるビットの数、すなわちビット密度を考え、各ディスクの位置を1ビット位置の許容差に維持しなければならないことを認識すると理解できる。このレベルの正確さの要件は、各DASDに対して先入れ先出し型の読出しバッファを設けることによって緩和することができる。バッファのサイズは、通常、単一セクタよりわずかに大きいものを扱えるように設定すると都合である。様々なDASDの間の小さなスキューは、データが各バッファから読み出されたときにデータを同期させることによって補正できる。この手法を用いると、スキューの問題が解決され、DASDの回転位置の許容差が緩和されるが、その代りに遅延が増加する。これは、データを使用するとき、最後のディスクがアドレスされたデータを読出しヘッドの下にもってくるまで、システムが待たなければならないためである。通常の状態では、この遅延は、3600rpmで動作しているDASDで通常見られる回転待ち時間より小さい程度の大きさである。バッファ誘発性遅延は、ディスクが数ビット位置移動するのに必要な時間である。実際の遅延は、ディスクが通常のDASDの約3倍の高速で回転していることによってさらに短縮される。

【0083】図10で、各DASD100nは、速度/位置制御機構90をもつ。これは、図1のアレイ/クラスタ制御装置5内に位置する。各速度/位置制御機構90は、制御ループ95をもつ。このループは、DASDからのインデックス・パルスの位置を、加算回路91内のすべての制御機構90に対して確定された基準と比較する働きをする。通常の場合、基準インデックス・パルスは、アレイ/クラスタ制御装置5内で水晶制御発振器92によって発生され、記憶サブシステム及び関連データ処理システムから見てインデックス・パルスの最適な出現時間を表す。

【0084】インデックス検出器93は、ディスク94上のインデックス・マークが通過することにインデックス・パルスを供給する。これは、DASD100n内のディスクの速度及び位置の尺度を提供する。インデックス検出器93からのインデックス・パルス及び水晶基準発振器92からの基準パルスが、加算回路91で比較され、その結果生じる誤差信号が、制御ループ95に送られる。好ましい実施例では、デジタル技術を利用して制御ループ95を実施するが、制御ループ95の実際の

実施態様はアナログ、ディジタル、またはハイブリッド技術のいずれでもよい。それが、DASDスピンドル・ドライブ・モータ97のコイルを励起する、DASDモータ・ドライブ回路96への信号を発生する働きをすると言うだけで十分である。制御ループは、DASD100nの速度を、水晶基準発振器92によって定義される速度に正確に合わせ、インデックス・パルスが正確な時間に丁度インデックス検出器を通過するようにDASD100nのディスクを位置決めすることができる。これによって、すべてのDASDの間の位置同期が実現される。あるDASDのディスク位置が他のDASDのディスク位置よりわずかに後ろにある場合、制御ループ95は、遅れたDASDの速度を、それが正しい速度で回転していても、適切な位置に達するまでわずかに加速させる。適切な位置に達した時点で、他のDASDと適合するように速度を下げ、適切な位置を維持する。

【0085】アレイ／クラスタ制御装置5に使用される電子回路のより包括的な図が、図11に示されている。中央演算処理装置との交信は、シリアル・インタフェース6を介して行なわれる。このインタフェースは、アレイ／クラスタ制御装置5内にあるシリアル・インタフェース・ドライバ／レシーバ・チップ120につながっている。ディジタル制御チップ122は、シリアル・インタフェース制御部123を含む。この制御部は、アレイ／クラスタ制御装置5とデータ処理システム2の間のシリアル・インタフェースを制御するためのものである。シリアル・インタフェース制御部は、データ処理システム2とアレイ／クラスタ制御装置5の間でやり取りされるすべてのデータ及び制御信号の適切な同期、緩衝及び復号を実行する。

【0086】サーボ・デモジュレータ・チップ127は、読出し線127aからディスク上に書き込まれたアナログ・サーボ・データを読み出し、位置のディジタル表示をバス127b上に供給する。

【0087】ディジタル制御チップ122はまた、ディジタル信号処理部125を含む。この信号処理部は、サーボ・デモジュレータ・チップ127の出力を表すバス127b上の読出しヘッド位置データに応答する。バス127b上の信号が処理されると、ボイス・コイル・ドライバ・チップ130に制御信号が供給される。ドライバ・チップ130は、DASDのヘッド・アーム・アセンブリと連動するアクチュエータのボイス・コイル型モータのコイルを励起する。サーボ・デモジュレータ・チップ127、ディジタル制御チップ122のディジタル信号処理部125、及びボイス・コイル・ドライバ・チップ130は、読出し／書込みヘッドをアドレスされたトラックの上方で位置決めするためにヘッド・アーム・アセンブリが移動され、データがトラックに書き込まれトラックから読み出されるのに適切な位置に維持されるように、ボイス・コイル・ドライバ・コイルを励起す

る。

【0088】さらに、サーボ・デモジュレータ・チップ127は、インデックス検出器の機能を果たし、インデックス信号をディジタル信号処理部125に渡す。インデックス信号は、先に図10に関して説明したように水晶基準発振器で比較され、適切な駆動信号がスピンドル・モータ・ドライバ・チップ96に送られる。ドライバ・チップ96から、図10に関して説明したように適切な駆動信号が発生され、スピンドル駆動モータ・コイル97に送られる。

【0089】ディジタル制御チップ122はさらに、符号化／復号部126を含む。この部分は、DASDディスクから読み出されたデータを復号し、そのデータをDASDディスク上に記録させる働きをする。記録されるデータは、アナログ波形コンディショナ・チップ124に渡される。コンディショナ・チップ124は、ディジタル波形を取り、それを適当なアナログ波形に変換して、マルチプレクサ・チップ128を介して読出し／書込みヘッド128a、128b、...、128nのうちの選択された1つのヘッドに送る。DASDディスクからデータを読み出すとき、ヘッド128a、128b、...、128nのうちの1つからアナログ信号が選択され、マルチプレクサ・チップ128によって増幅され、アナログ波形及びクロック生成機構チップ124に渡され、そこで調整されディジタル化される。位相ロック・クロック信号が、読み出されたデータから取り出され、ディジタル化された読出しデータと共にディジタル制御チップ122の符号化／復号部126に供給される。次に符号化／復号部126は、それをディジタル波形に変換し、ディジタル制御装置チップ122のシリアル・インタフェース制御部123へ転送する。

【0090】本発明の電気的態様及びシステム態様について説明したので、次に図12図13及び図14に移る。図12は、記憶サブシステム内で使用される1つのDASDの平面図である。図13及び図14は、図12に示したDASDの線X I I-X I Iに沿って切りとった断面図である。可能な場合、説明される要素を両方の図に示し、同じ参照番号で識別する。ベース・プレート150とカバー151は、組み合わさって、出来た格納機構内に位置するDASDの各種構成要素のシールを行なう。

【0091】公称直径65mmの複数のディスク160a、160b、160c、160dが、スピンドル・アセンブリ162上に回転できるように配置され、その一端はベース・プレート150で、他端はカバー151で支持されている。スピンドル・アセンブリ162の一部として、スピンドル駆動モータ164、及びディスク160a、160b、160c、160dを適切な位置に保持するためのスペーサ213とクランプ212が組み込まれている。スピンドル・アセンブリ162につい

25

ては、後で図17を参照して詳しく説明する。

【0092】ディスク160a-160dの各表面は、データを記録するために適当な磁性記録材料でコートされている。図16に示したように、複数の磁気トランスジューサ170a及び170a'ないし170d及び170d'が、データを読み出し書き込むため、ディスク160aないし160dの上下のコートされた表面と共同作用するように位置決めされている。記録ヘッド170は、当該のサスペンション・アセンブリ171a及び171a'ないし171d及び171d'の端部に装着されている。ヘッド170のサスペンションは、通常のジンバル設計に従い、システムの動作中にヘッドをディスク表面から約6.5マイクロインチ(0.165ミクロン)浮かせることができる。

【0093】再び図12を参照すると、各DASDアセンブリは、DASDへのすべての必要な電力線及び信号線を担持するコネクタ169を含む。

【0094】図15を参照すると、各ヘッド/サスペンション・アセンブリは、コーム・アセンブリ173の対応するアーム173a-173eに固定されている。コーム・アセンブリ173は、ベアリング・カートリッジ175の周りを回転するように配置されている。ベアリング・シャフト端部175a及び175bは、ベアリング175を両端支持するように、ベース・プレート150及びカバー151を貫通してそれに結合し、カバー151をベース・プレート150と係合した状態に保持する助けをする。駆動コイル176から構成されるアクチュエータ・コイルは、コーム・アセンブリ173の、アーム173a-173eと対向する端部に装着されている。

【0095】コイル176がいずれかの極性の直流によって励起されると、駆動コイル176内の電流が、それぞれカバー151とベース150に固定された永久磁石アセンブリ177aと177bの磁界と共同作用して、コーム・アセンブリ173をベアリング175の周りで回転させ、ヘッド170をディスク160上の所望のトラックの上方で位置決めし、データの読出し/書き込み中にそれらのヘッドをトラック上方の正しい位置に維持する。磁石アセンブリ177aと177bはそれぞれ1対の永久磁石177a1と177a2、177b1と177b2を含み、各対はそれぞれ軟磁性体177a3と177a4のベース部材上に装着されている。これらのベース部材は、駆動コイル176から見た磁気回路を完成し、磁気アクチュエータの駆動効率を向上する低磁気抵抗経路を提供する。1対のクラッシュ・ストップ178a及び178bは、それぞれその中点にエラストマ・クッション178cをもち、コームが許容範囲を超えて移動するのを防止する働きをする。フレキシブル・ケーブル・アセンブリ179は、磁気ヘッド170及び駆動コイル176への必要な接続を行なう。

26

【0096】図17に示したスピンドル・アセンブリ162は、シャフト190を含み、その両端は、ベース・プレート150及びカバー151中に設けられた対応する穴と対合し、シャフトを両端で支持するようにそれらに結合されている。シャフトを両端で支持することによって、いわゆる「タワー」装着に付随する傾斜及び振動の問題は完全に回避され、完成したアセンブリは振動及び共鳴がほとんどない。シャフト190に取り付けられた1対のベアリング191と192は、ディスク駆動モータの極193を支持し、それを固定されたシャフト190の周りで回転させる。

【0097】シャフト190に取り付けられた1組の巻線194が駆動電流によって励起され、回転磁界を発生する。この回転磁界は、鋼鉄の極193の内面に装着された永久磁石200と共同作用して、その極を回転させる。鋼鉄のスペーサ201はベアリング192と係合し、極193の下部を支持する。極193の外面に取り付けられたアルミニウム・ハブ210は、その外径がディスク220の内径より小さく、熱膨張の余地を与える。ディスク220及びスペーサ213は、シャフト190の周りで心合せされ、クランプ212によってハブ210上で定位置に保持されている。

【0098】図18は、4台のDASD230a、230b、230c、230dが単一のプラグ接続可能なカード231上にどのように装着されるかを示す。すなわち、このような各カードは、単一のアレイを構成し、複数のこのようなカードがクラスタを形成する。このカードは、通常のプリント回路素材から作成され、接触舌232から個々のDASDに延びるランド(図示せず)を含む。個々のDASD230はそれぞれ、プラグ235a-235dを介してカードに接続される。

【0099】図19は、DASD230がどのようにカード231に固定されるかを示す。各DASD位置は、キーホール・スロット238a、238b、及び第3のスロット(図示せず)を含む。これらのスロットは、DASD230上の対応するショック・マウント240a、240b、及び第3のマウント(図示せず)、それらに関連するプラグ234b1、234b2、及び第3のプラグ(図示せず)と一列に並び、

【0100】図20は、DASDプラグ及びショック・マウントの配置をより詳細に示す。DASDはプリント回路カード231と係合して保持されるが、ショック・マウントによってカードに関連するショックから機械的に隔離される。各DASDは、3つのショック・マウントを含み、それぞれが真中を通るフェール・ピン242、及びエラストマ・グロメット240をベース150に保持するネジ部分243をもつ。ショック・マウント型グロメットは、通常の形状をもち、プリント回路カードの穴の肩を収容する環状のスロットを周囲に含む。グロメットをカード内の変形スロットの大きい部分に差し



込むことによって、DASDをカード231上に装着する。次にDASDを移動させて、グロメット・スロットを鍵形スロットの狭い部分に押し込む。プラグ234を鍵形スロットの大きい部分に差し込むことによって、DASDはスロットの狭い部分に保持される。

【0101】いま説明したDASDの機械的装着は、比較的簡単であり、使いやすい配置を提供するだけでなく、DASDとカードの間の多数の電気的接触に関する要件を満たす。システムのプラグ接続可能な各構成要素間の機械的隔離と電気的接触というこの二重の要件は、従来は一般に、固定されたショック・マウントとプラグ接続可能なケーブルによって満たされてきた。この手法は、ほとんどのパーソナル・コンピュータ・システムでいわゆるハード・ドライブの装着の際にとられている。この二重要件のため、たとえば機械的装着の完了後に必要なケーブル接続ができないなど、エラーの機会が増大する。これは、故障を直すためにコンピュータを部分的に分解しなければならないことを意味する。さらに、ケーブルは配線を誤りやすく、システムのコストを増加させる。

【0102】図20に示すように、DASD230のプラグ接続可能なコネクタ169は、ソケット241につながるアパーチャと隣接するが、それから離れた位置にくるように構成されている。DASDを移動させて、ショック・マウントをキーホール・スロットの狭い部分に位置決めすると、コネクタ169は、コネクタ169上のランドがソケット241内の対応する接点と接触するように、ソケット241のアパーチャ内に入る。従来のソケットは、DASDとカード231の間の機械的隔離をもたらさないので、新規な配置が提供される。コネクタ169のランドとカード231上の対応するランド、たとえばランド245及び256との間の電気的接続は、パネ247及び248によって実施される。挿入中のソケットの位置決めは、2本のピン260によって行なわれる。これらのピンは保持用座金260aを含み、これらの座金は、カード231の過大なサイズの穴中でピンを浮かせる。これは、DASDコネクタ169の挿入中の過剰な移動を防止するが、その後係合が外れて、DASD230をカード231から機械的に隔離する。

【0103】本発明の利点は、交換の前にシステムの電源を遮断し、交換後にシステムの電力を再度投入する必要なしに、故障したDASDを取り外し、それを動作可能な装置と交換できることである。したがって、この動作モードは、「ホット・プラグング」と呼ばれる。図21で、ホット・プラグング機能の機械的態様を図示する。各DASDアレイ300は、信号及び電源用の複数の接触ランドを含むプラグ接触部分302をもつカード301を備えている。プラグ接触部分302上のランドは、データ処理システムのフレームに取り付けられたソケット310の対応する接点と共同作用する。DASD

がソケット310で電源にプラグ接続できるように、各ランド320は電源ランド321及び322よりも物理的に短くなっている。このため、DASD300内の回路及びアレイ/クラスタ制御装置5内の回路がプラグ接続操作中に優先的な信号を受け取らないことが保証される。信号回路と電源回路が混合方式でまたは同時に接続される場合、それらの回路はスプリアス信号を発生し、アレイ/クラスタ制御装置に、またはさらに悪くするとデータ処理システムにエラーをもたらす。信号接続が行なわれる前にDASD300の回路に適切に電力を供給することによって、スプリアス信号の発生が回避される。プラグ302上のランドの機械的配置は、電源ランドが接続されるまで信号ランドへの接触を物理的に防止する。挿入中、電源接触を行なってから、信号接触を行なうまでの間の時間遅延により、すべての回路に適切に電力が供給され、それによってエラー信号の発生が防止される。

【0104】

【発明の効果】

20 【0105】本発明によれば、単一ドライブ・システムよりも信頼性の高い多重ドライブ・ディスク記憶システムが得られる。

【0106】本発明によれば、多重ドライブ・ディスク記憶システムを電力遮断する必要なく、かつデータの損失なく、個々のドライブを接続及び切断できる、多重ドライブ・ディスク記憶システムが得られる。

【0107】本発明によれば、4つのドライブをもち、そのうちの1つのドライブが故障してもデータの損失を起こさない、または実質的に性能を損なわない、ディスク記憶システムが得られる。

【0108】本発明によれば、データの並列検索によってデータがより高速で検索できるように、高速、低コストのアクチュエータを使用して比較的により少量のデータにアクセスすることのできる、ディスク記憶システムが得られる。

【0109】本発明によれば、アクチュエータ速度の損失またはシステム信頼性の低下を受けずに、各アクチュエータによってアクセスできるデータの量を減らす、ディスク記憶システムが得られる。

40 【図面の簡単な説明】

【図1】本発明のディスク記憶システムのアーキテクチャを示す機能ブロック図である。

【図2】データの変換、及びデータ処理システムから記憶サブシステムへのデータの転送を示す機能ブロック図である。

【図3】ディスク・ドライブが動作不能になった条件下での記憶サブシステムからデータ処理システムへのデータの転送を示す機能ブロック図である。

【図4】システム電源投入及びシステム・リセット中にアレイ/クラスタ制御装置が利用するプログラムを示す

データ流れ図である。

【図5】制御プログラムの、システムの電源を遮断せずに欠陥DASDの取外し及び正常DASDの代替使用を行なう部分を示すデータ流れ図である。

【図6】アレイ/クラスタ・プログラムの、欠陥があると判定されたDASDの代りに新しいDASD上にデータを再構築する部分を示すデータ流れ図である。

【図7】プログラムの、DASDから読み出されたデータを取り、パリティ検査し、元のデータを再生する部分を示すデータ流れ図である。

【図8】プログラムの、データ処理システムからのデータをDASDサブシステムに書き込む部分を示す左半分のデータ流れ図である。

【図9】プログラムの、データ処理システムからのデータをDASDサブシステムに書き込む部分を示す右半分のデータ流れ図である。

【図10】プログラムの、スピンドル駆動モータを制御するのに使用される部分を示すデータ流れ図である。

【図11】記憶サブシステム制御装置によって利用されるシステムの電子回路の概略図である。

【図12】記憶サブシステム制御装置で使用されるディスク・ドライブの平面図である。

【図13】図12に示したディスク・ドライブの側面図である。

【図14】図12に示したディスク・ドライブの側面図である。

【図15】アクチュエータ及びヘッド・アーム・アセンブリの展開図である。

【図16】ヘッド・アーム・アセンブリ及びディスクの両端の部分側面図である。

【図17】ベアリング支持体の中心で切ったスピンドル駆動モータの側面図である。

【図18】5.25インチのディスク・ドライブの標準の形状因子の整数分の1に適合し、個々のディスク・ドライブを切断及び接続することのできる、カード上に装着された4台のディスク・ドライブの配置図である。

【図19】個々のDASDがどのようにカード上に装着されるかを示す図である。

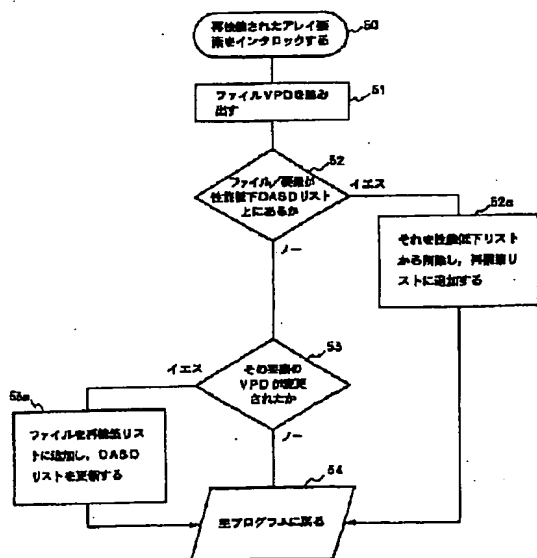
【図20】DASD用のショック・マウント及びソケットの詳細な断面図である。

【図21】個々のクラスタをアレイから外して交換できる、プラグ及びガイドの配置を示す図である。

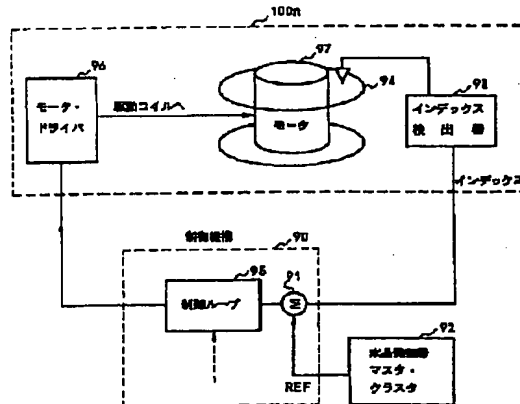
【符号の説明】

- 1 ディスク記憶システム
- 2 データ処理システム
- 3 中央演算処理装置
- 4 ランダム・アクセス記憶装置 (RAM)
- 5 アレイ/クラスタ制御装置
- 7 記憶装置アレイ
- 8 記憶装置アレイ
- 9 記憶装置アレイ
- 10 データ処理システム
- 101 データ・パッチ
- 105 パッファ及びECC生成機構
- 110 パリティ生成機構

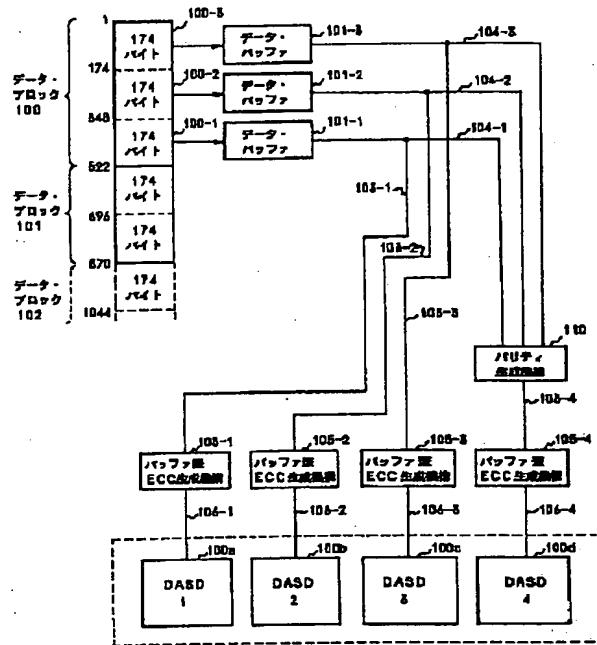
【図5】



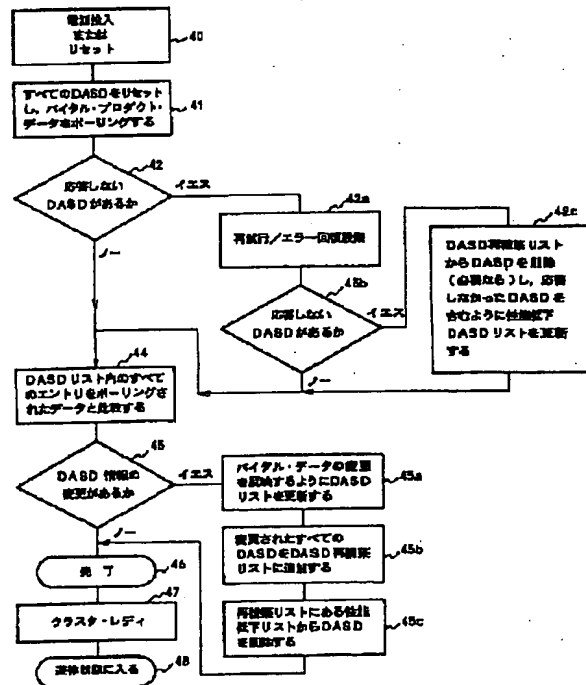
【図10】



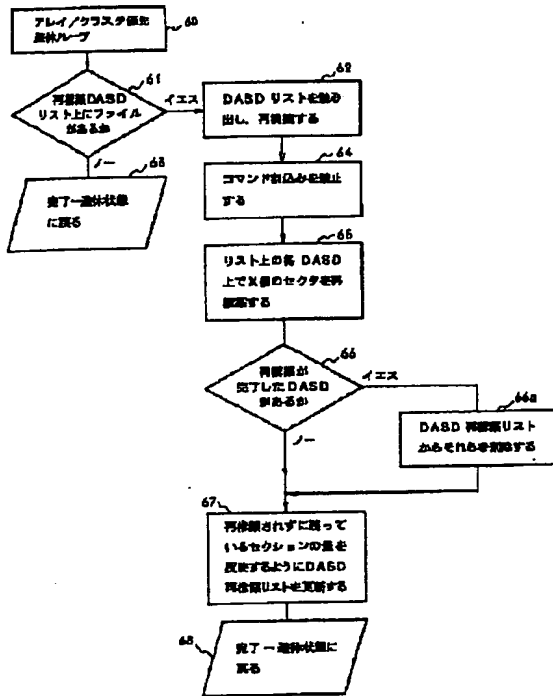
【圖 2】



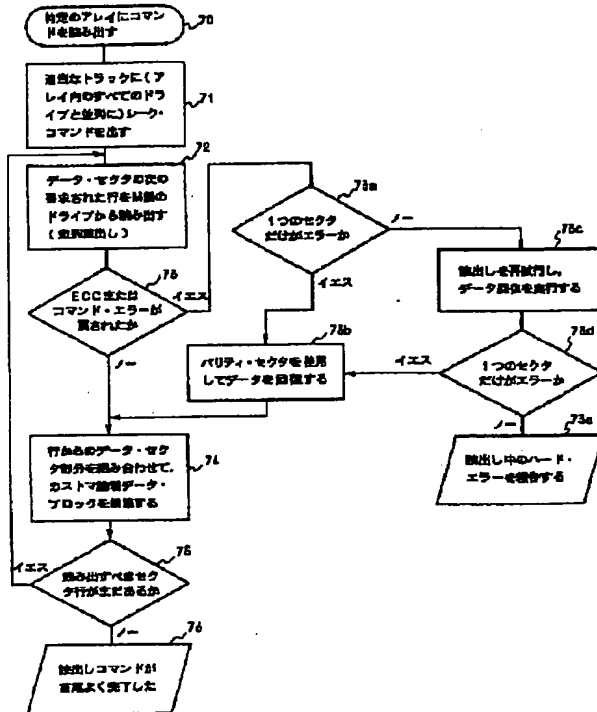
【图 3】



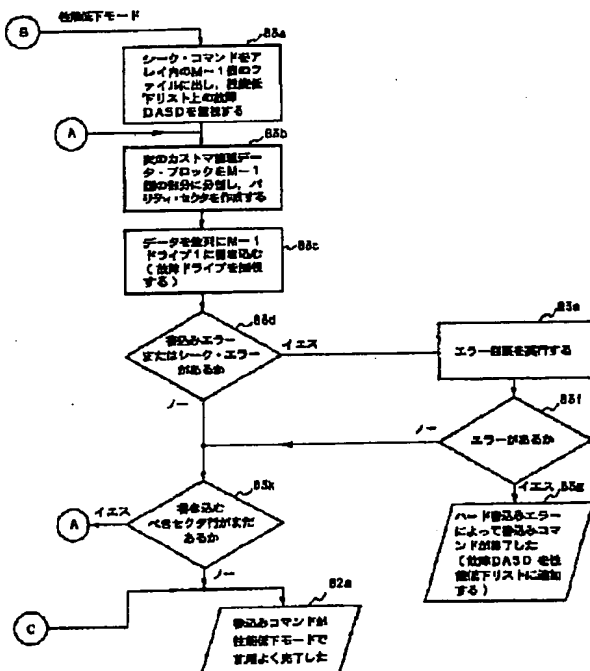
【図 6】



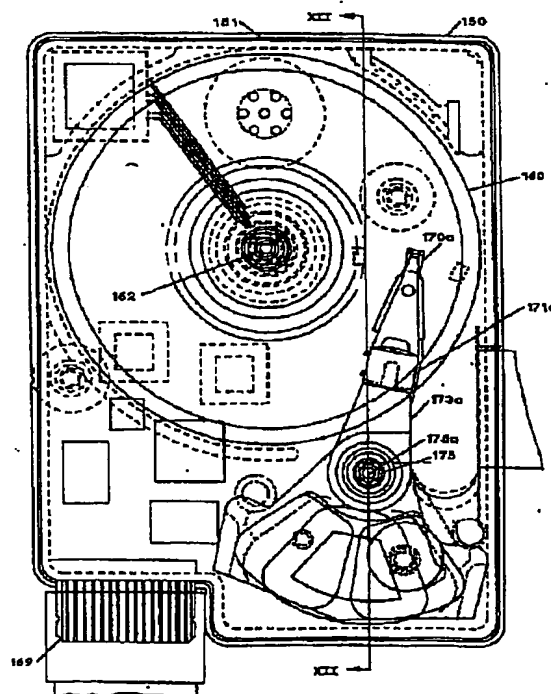
【図 7】



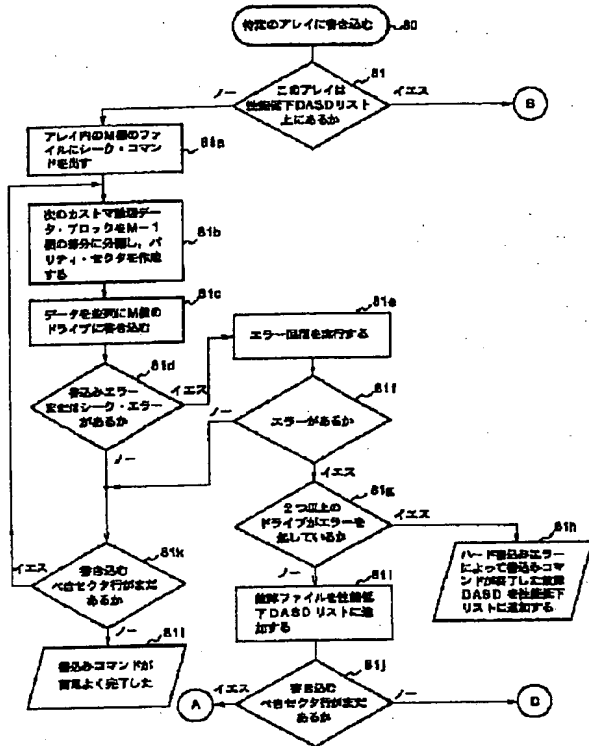
【例 9】



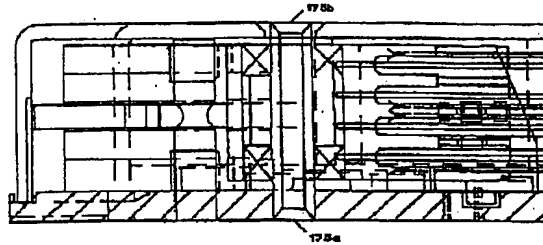
【例 12】



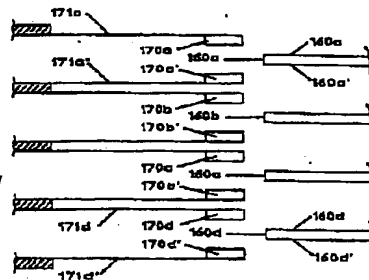
【図8】



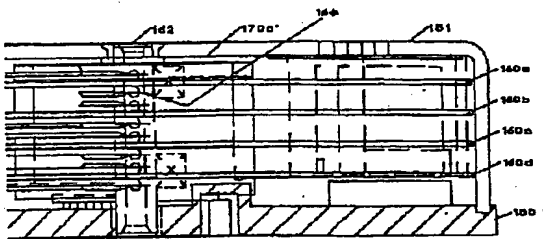
【図13】



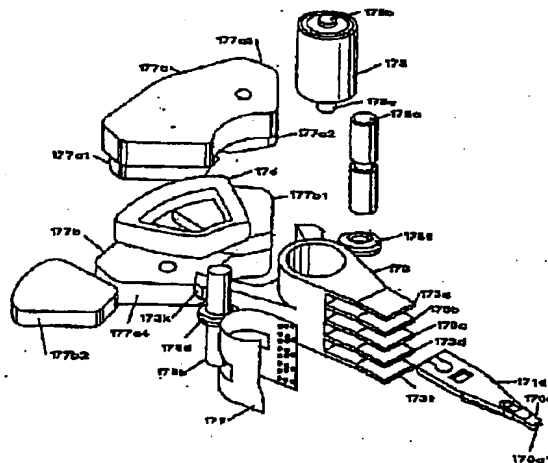
【図16】



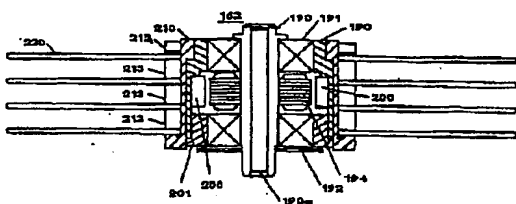
【図14】



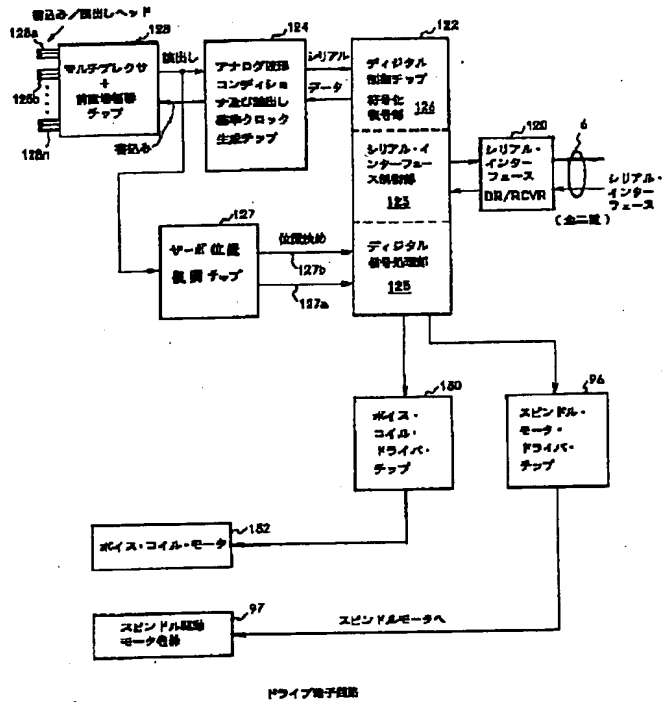
【図15】



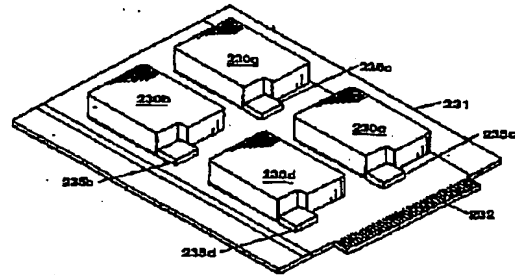
【図17】



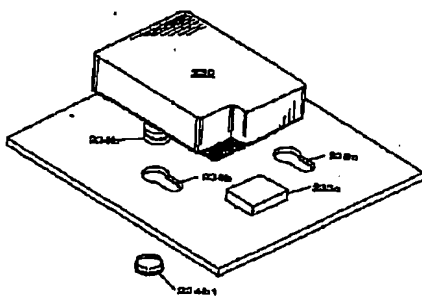
【図11】



【図18】



【図19】



【図20】

